College of the Holy Cross, Spring Semester, 2018
MONT 101N – Analyzing Environmental Data, Final Exam Solutions
Tuesday, May 15

I. *Briefly* identify *any 4* of the following. If you submit answers for more than 4, the best 4 will be counted.

(A) (5) The Central Limit Theorem

Solution: The Central Limit Theorem says that if independent and identically distributed values Y_i, $i = 1, \ldots, N$ are sampled from any distribution with finite mean μ and SD σ, then the distribution of the *sample means*

$$\overline{Y} = \frac{1}{N}(Y_1 + \cdots + Y_N)$$

becomes more and more normal as $N \to \infty$. Precisely, the distribution of

$$\frac{\sqrt{N}(\overline{Y} - \mu)}{\sigma}$$

approaches a standard normal distribution as $N \to \infty$.

(B) (5) A 95% confidence interval for a population mean

Solution: The 95% confidence interval (for a sample size $N > 30$) is the interval $\overline{Y} \pm 1.96\frac{S}{\sqrt{N}}$. The idea is that if we repeat the sampling process many times, we would expect about 95% of the intervals constructed this way to contain the corresponding population mean. If $N < 30$, then the factor 1.96 would be replaced by the .025 percentage point for the t-distribution with $N - 1$ degrees of freedom.

(C) (5) A t-distribution

Solution: When sampling from a normal population with mean μ, the t distribution describes the "standardized sample means" where the population σ is estimated by the sample S: $\dfrac{\overline{Y} - \mu}{S/\sqrt{N}}$. This gives the t-distribution with $N - 1$ degrees of freedom.

(D) (5) Standard normal distribution

Solution: The normal distribution with $\mu = 0$ and $\sigma = 1$.

(E) (5) The box-and-whisker plot of a data set.

Solution: A plot showing the minimum, first quartile, median, third quartile and maximum of a data set. The "box" is drawn between the first quartile and the third quartile; the whiskers are thinner lines extending to the left and right (or down and up) to the minimum and maximum, respectively.

(F) (5) The p-value of a test of hypotheses

Solution: The p-value is the *smallest value of the Type I error probability α for which the null hypothesis would be rejected with the observed data.* In more intuitive terms, it is the chance that the observed test statistic, or a "more extreme" value, would be observed from a sample *if the null hypothesis is true.* The smaller the p-value is, the stronger the evidence is for rejecting the null hypothesis.

II. Short answer

(A) (5) If you want to get polling results with comparable accuracy, do you need a bigger sample size if the polling is done in Texas than you do if the polling is done in Delaware? Explain by using formula for the standard error for a percentage.

1. *Solution:* The answer is NO. The formula standard error (SE) for the percent shows why. The SE for the percent is always bounded above by $\sqrt{\frac{(.5)(.5)}{N}}$ – the SE for the percent depends only on the sample size, not on the size of the whole population. (This *is assuming* that the size of the population is enough larger than the size of the sample that the difference between sampling with replacement and sampling without replacement can be ignored, but that would be true for any population the size of a U.S. state.)

(B) (5) True/False and explain: If the p-value of a hypothesis test works out to $p = .034$ then we can say that the null hypothesis was *definitely false.*

Solution: This is FALSE. The p-value gives the probability of getting the observed test statistic, or something even more extreme, under the assumption the null hypothesis is true. We cannot say the null hypothesis is definitely false even with a p-value as small as $p = .034$.

III. Let X represent a normal random variable with $\mu = 4.3$ and $\sigma = 1.4$.

(A) (10) Find $P(3.2 < X < 5.0)$

Solution: This probability is the same as

$$P\left(\frac{3.2-4.3}{1.4} < Z < \frac{5.0-4.3}{1.4}\right) = P(-0.79 < Z < 0.5) = .2852 + .1915 = .4767$$

(B) (10) Suppose that values of X are sampled $n = 10$ times independently. Give the *formula for computing* the probability that exactly 5 out of the 10 values satisfy $3.2 < X < 5.0$. (The number the correct formula gives is roughly .2434 if you want to check your answer by computing the value.)

Solution: This is a binomial experiment. The probability that exactly 5 out of the 10 values are in the range is

$$\binom{10}{5}(.4767)^5(1-.4767)^5 = 252(.4767)^5(.5233)^5 \doteq .2434$$

IV. A Gallup poll released on April 3, 2017 surveyed a random sample of 706 adults nationwide. It reported "that 59% of those sampled say the environment should be prioritized over energy production."

(A) (20) Find the 95% confidence interval for the actual percentage of Americans who feel this way. (Use the conservative method for estimating the standard error.)

Solution: The confidence interval is

$$59\% \pm 2(\frac{\frac{1}{2}}{\sqrt{706}} \times 100\%) = 59\% \pm 3.8\%$$

This extends from 55.2% to 62.8%

(B) (5) Based on your answer to part (A), is 55% a "believable" value for the percentage of all Americans who feel that way?

Solution: Probably not. The lower endpoint of the confidence interval is $55.2 > 55$, so 55 lies outside the most probable range of values for the population percentage.

V. (25) A study was conducted by the Florida Game and Fish Commission to assess the amounts of chemical residues found in the brain tissue of brown pelicans. In a test for the pesticide DDT, random samples of $N_1 = 10$ juveniles (young birds foraging independently) and $N_2 = 13$ nestlings (birds still being fed by their parents) produced the following data (DDT measurements in parts per million). Test the alternative hypothesis that mean amounts of DDT found in juveniles and nestlings differ versus the null hypothesis that they do not differ. Use $\alpha = 0.05$.

Juveniles	Nestlings
$N_1 = 10$	$N_2 = 13$
$\overline{Y}_1 = .041$	$\overline{Y}_2 = .026$
$S_1 = .017$	$S_2 = .006$

(Notes: This data has important implications regarding the accumulation of chemical residues over time. You may assume that the population SD's are equal to the sample SD's in deciding how to set up the test.)

Solution: Since $N_1 = 10$ and $N_2 = 13$ are both less than 30, we want to use the "small sample" test for equality of population means from the fifth line in our summary table about hypothesis tests. We have

- $H_0 : \mu_1 = \mu_2$
- $H_a : \mu_1 \neq \mu_2$ (a two-tail test, indicated by the way the question is phrased: " Test the alternative hypothesis that mean amounts of DDT found in juveniles and nestlings differ versus the null hypothesis that they do not differ"

The test statistic will have a $t(9)$ distribution (the number of degrees of freedom is $9 =$ the smaller of $10-1$ and $13-1$), so the rejection region for the two-tail test will be $t < -2.26126$ and $t > 2.26126$ (from the .025 column since the $\alpha = .05$ is split even over the upper and lower tails).

We compute

$$t = \frac{.041 - .026}{\sqrt{\frac{(.017)^2}{10} + \frac{(.006)^2}{13}}} \doteq 2.665$$

This is in the rejection region, so we reject the null hypothesis that there is no difference in the DDT levels in the brains of the juvenile birds and the nestlings.